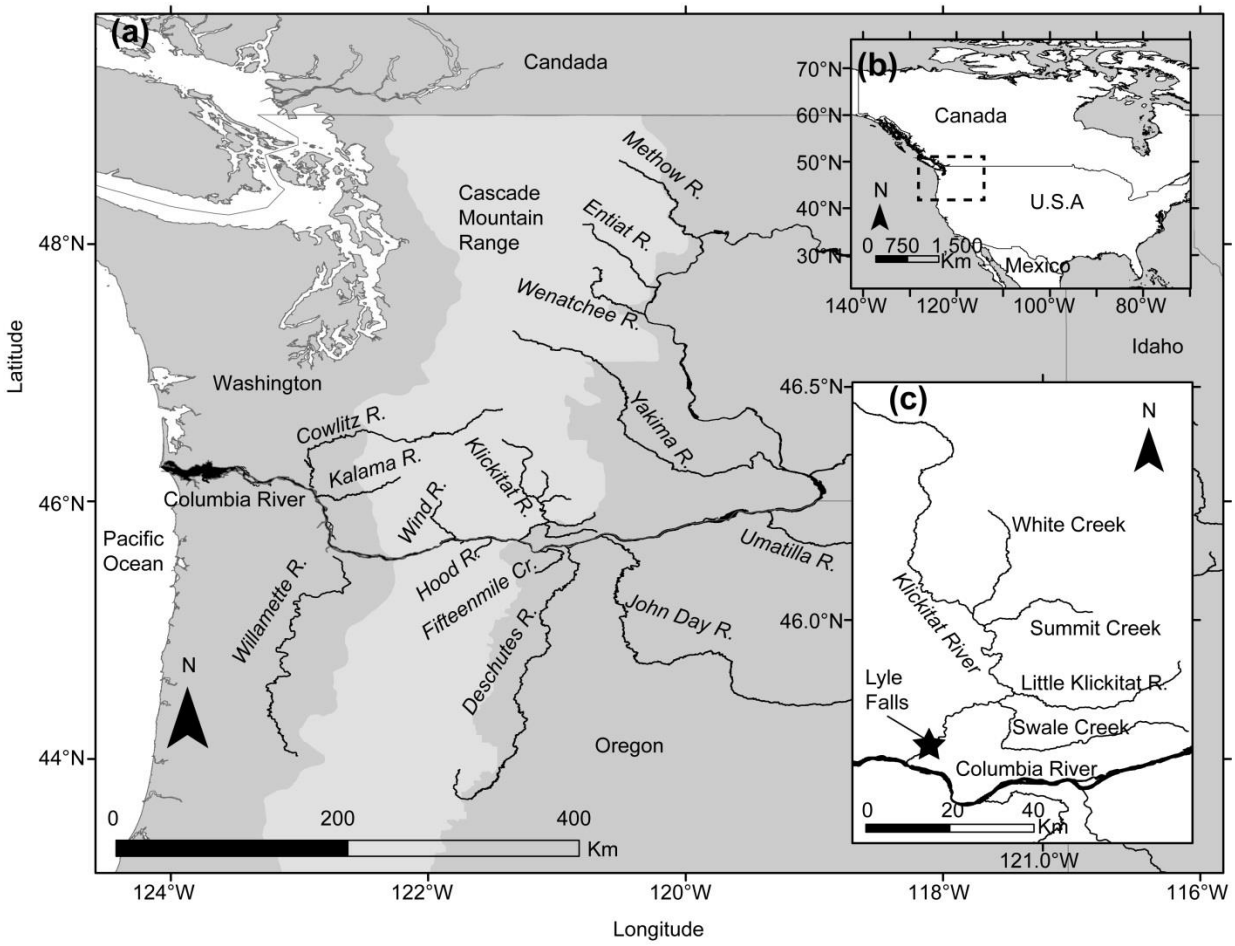
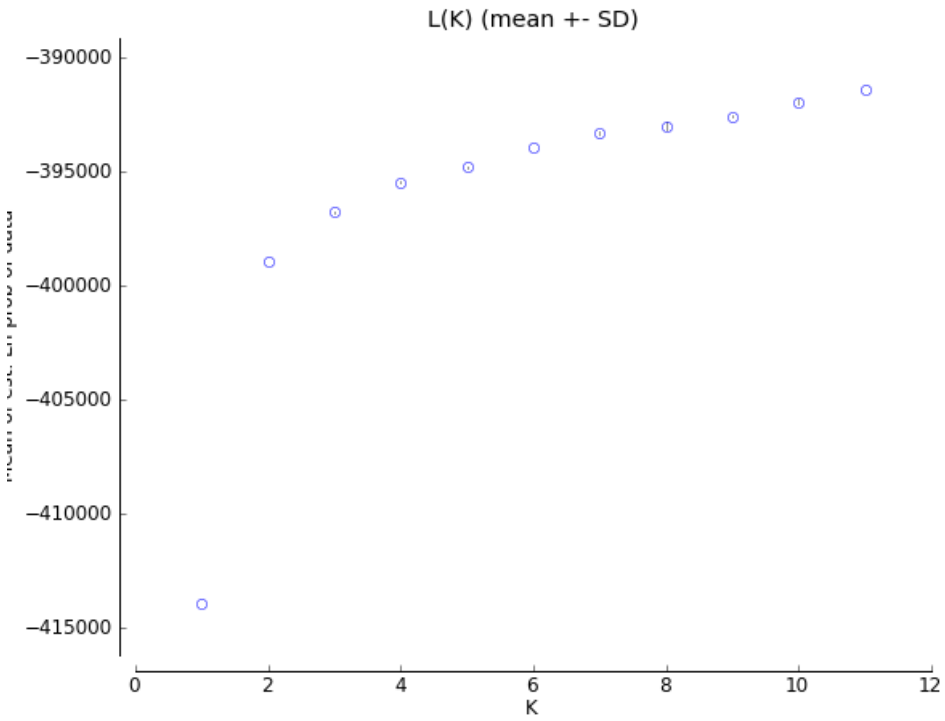
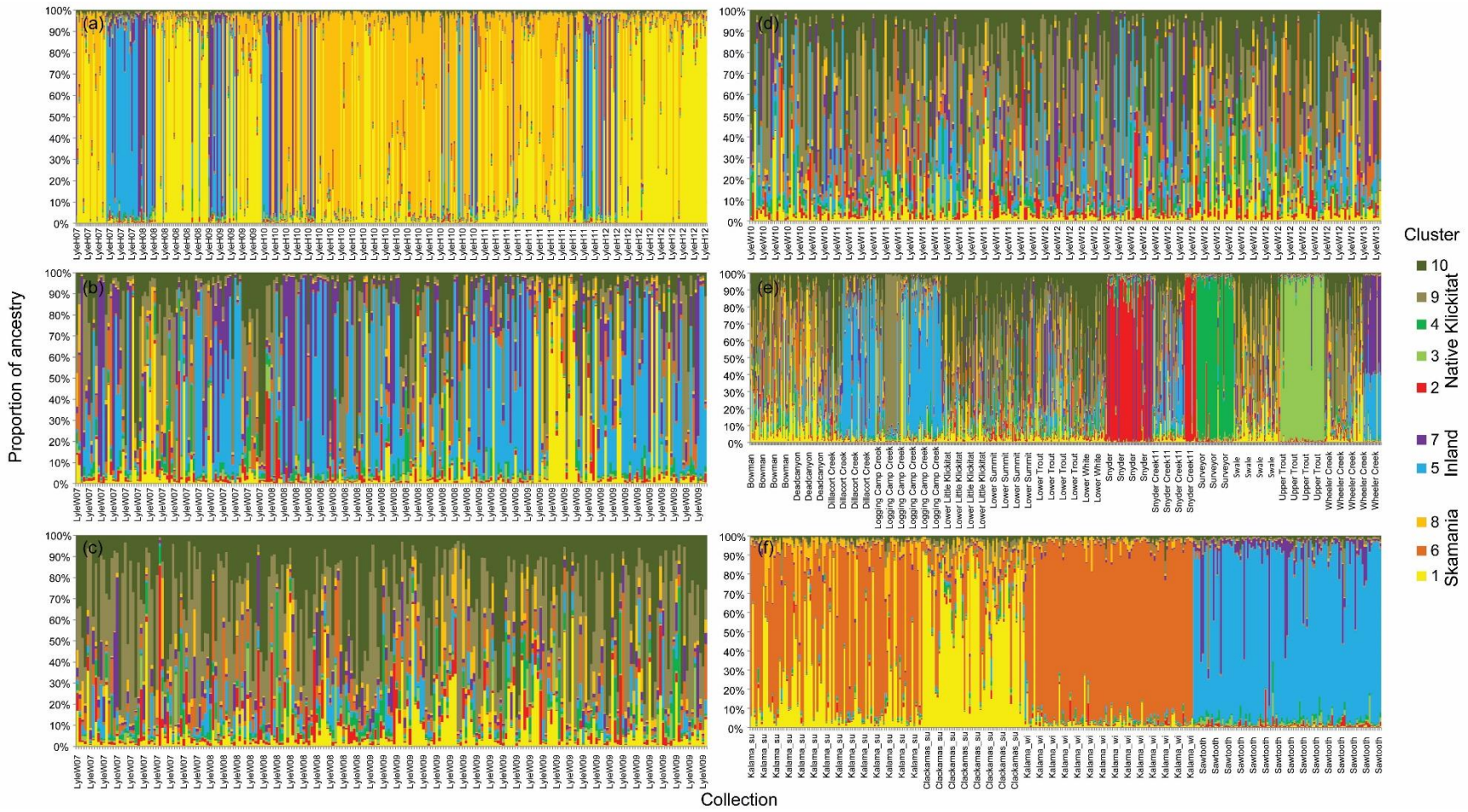


Supplemental Material Figure S1. Map of the study site. (a) The Klickitat River tributary and neighboring tributaries of the Columbia River Basin. (b) North America with a dashed box to indicate the study region in the Columbia River Basin. (c) The Klickitat River with labels on secondary tributaries within the subbasin and the site where steelhead were collected at Lyle Falls.



Supplemental Material Figure S2. Structure results for the mean of LPr(K) versus values of K (1–11) for the pre-screening of the sample used for analysis. See Supplemental Methods S1.



Supplemental Material Figure S3. Structure analysis to pre-screen the sample. The analysis included (a) hatchery-origin and (b-d) natural-origin steelhead adults collected from Lyle Falls in the Klickitat River between 2007–2013, (e) natural-origin juvenile collections from in-stream surveys throughout the Klickitat, and (f) hatchery-origin collections of fish that are spawned in the Skamania Hatchery outside the Klickitat River and annually released in-basin as juvenile smolts ("Kalama_su" and "Clackamas_su",

representing Skamania stock), a natural-origin winter-run collection from the lower Columbia River ("Kalama_wi"), and a hatchery-origin collection representing the inland lineage of steelhead (Sawtooth Hatchery). A set of 237 individuals (c) were found to meet two main criteria for inclusion in this study: 1) natural-origin and 2) members of the native Klickitat River population of steelhead. This STRUCTURE analysis allowed us to identify and exclude any unmarked hatchery-origin steelhead ("Skamania") as well as out-of-basin stray fish ("Inland"). I.e., we excluded individuals having >80% membership to clusters 1, 6, and 8 and clusters 5 and 7 for Skamania and Inland, respectively. Individual membership to these 10 Structure clusters was used as a covariate in the GLM and MLM univariate analyses to account for population structure.

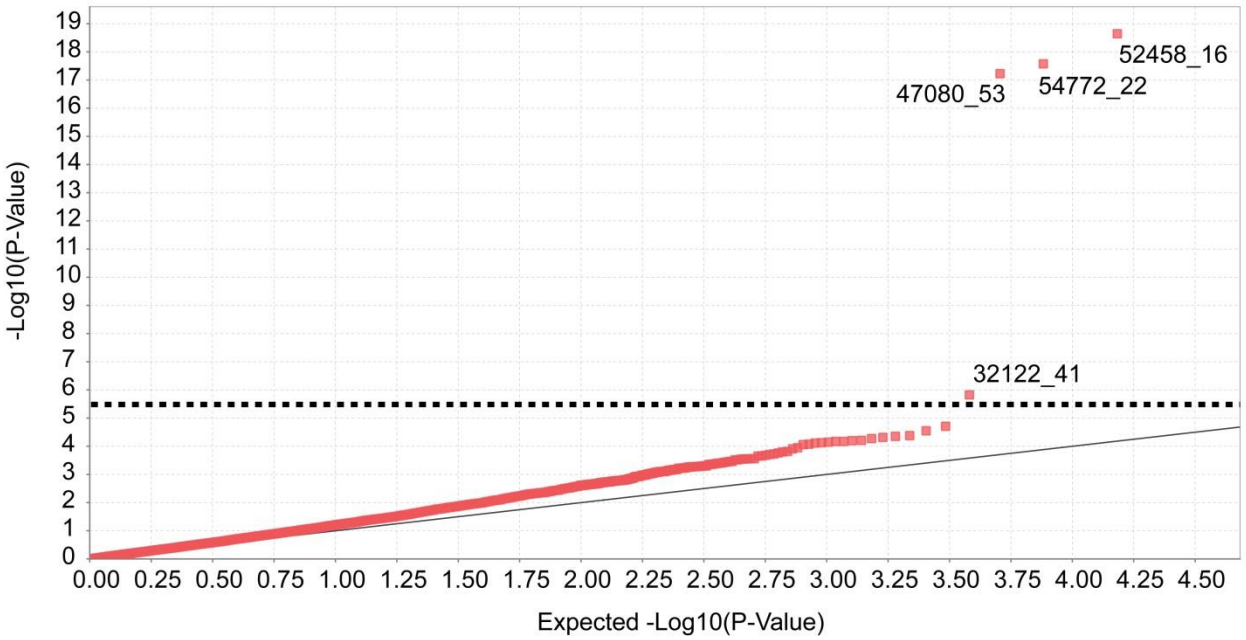


Figure S4. QQ plots showing the Expected -Log10(P-Value) vs. -Log10(P-Value) for a general linear model and of the migration timing trait. The heavy dashed lines indicate the Bonferroni corrected alpha level of 0.05.

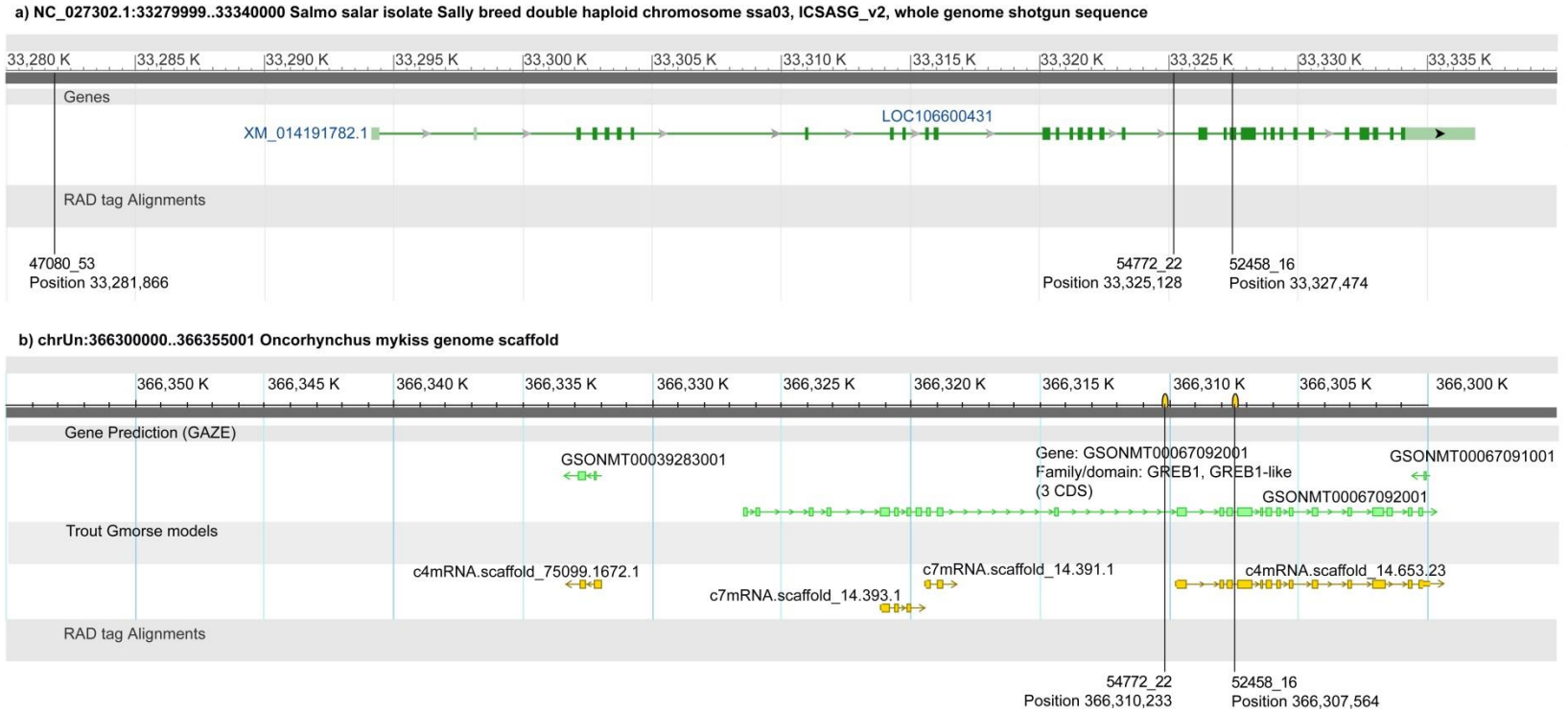


Figure S5. (a) A graphical representation of locations of the 3 candidate SNPs in the S. salar chromosome ssa03, and (b) two of these candidate SNPs 52458_16 and 54772_22 in the *O. mykiss* unknown chromosome scaffold (chrUn). The region between the SNPs contains 3 CDS of a predicted gene in the GREB1/GREB1-like family which is corroborated across these annotated genomes.

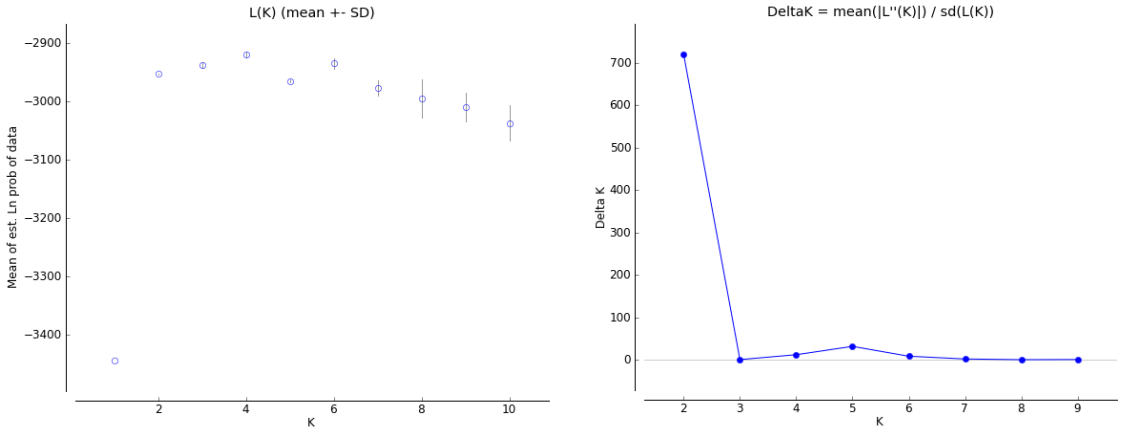


Fig. S6a

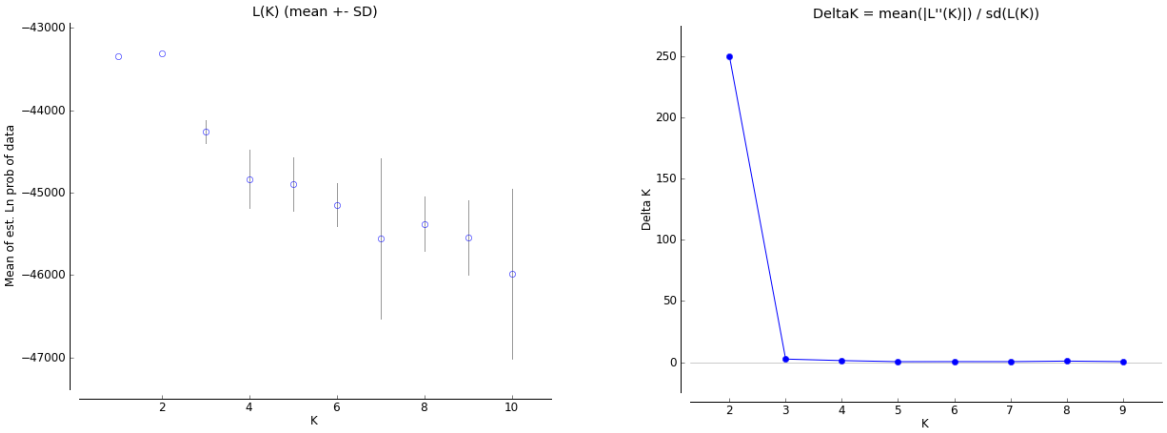


Fig. S6b

Supplemental Figure S6. Structure results (LPr(K) and Evanno et al's (2005) Delta K) that support K=2 for the two marker sets, (a) 18 candidate SNPs associated with migration timing and (b) 180 SNP markers used for genetic stock identification of Columbia River steelhead.

Supplemental Methods S1: Methods/Results for pre-screening of sample for analysis

We pre-screened a sample of hatchery-origin (N=458) and natural-origin (N=805) steelhead adults collected from Lyle Falls between 2007–2013 and utilized a subsample of individuals that were found to meet two main criteria for inclusion in this study: 1) natural-origin and 2) members of the native Klickitat River population of steelhead. While unmarked individuals (i.e., lacking an adipose fin-clip) can be putatively classified as natural-origin, it was necessary to perform a STRUCTURE v.2.3.4 (Pritchard et al. 2000, Fig. S2) analysis to identify and exclude any unmarked hatchery-origin steelhead as well as out-of-basin stray fish. Therefore we analyzed a dataset of the adult steelhead from Lyle Falls using the published 180 SNP marker set, and included the following 3 groups of reference populations: 1) natural-origin juvenile collections from in-stream surveys throughout the Klickitat (native Klickitat River stock; Narum et al. 2006, 2011), 2) hatchery-origin collections of fish that are spawned in the Washougal Hatchery outside the Klickitat River and annually released in-basin as juvenile smolts (Skamania stock), and 3) a hatchery-origin collection (Sawtooth Hatchery stock) that represented the inland lineage (originating upstream of the Klickitat River) which has been a source of temporary and permanent straying to the Klickitat River. STRUCTURE analyses were performed using 5 runs for each value of K (1–11). We chose K=10 in order to generate individual Q values for subsequent GWAS univariate analyses, then we ran 40 runs of K=10, selected the top 25% (based on estimated LnP(K)) of runs, and averaged across these top runs using CLUMPP. CLUMPP was run with the "Greedy" option (Jakobsson & Rosenberg 2007). All natural origin steelhead captured at Lyle Falls were screened for pure strays (assigning with >80% probability to the STRUCTURE clusters representing either the Skamania or inland

hatchery reference populations, Fig. S3). From the remaining natural-origin steelhead we chose 320 individuals, however variance in genotype quality (filtered in the steps below) reduced our dataset to 237 individuals. These individuals represent three migration timing groups, 3 sample years (2007–2009), and both male and female gender (Table S3).

For the pre-screen STRUCTURE analyses evaluating K values 1–11 using the panel of 180 non-candidate SNPs, we observed the most dramatic increase in mean LnP(K) between K=1 and K=2 (Fig. S2). This pattern was expected and indicative of the split between two major lineages (coastal and inland) that are present in the Columbia River Basin as described by Blankenship et al. (2011) and Matala et al. (2014). The mean LnP(K) continued to rise steadily across all K values suggesting multiple populations represented among the Klickitat River steelhead and reference collections (Fig. S2). To be conservative and account for as much population structure as possible, we chose K=10 in order to generate individual Q values for subsequent GWAS univariate analyses (Repeating these univariate analyses using a mid-range K value of 6 produced nearly identical results). CLUMPP was used to identify the best configuration across multiple STRUCTURE runs K=10, which generated a high pairwise similarity score (H' =0.92) owing to high consistency across the STRUCTURE runs.

References

Blankenship SM, et al. 2011 Major lineages and metapopulations in Columbia River *Oncorhynchus mykiss* are structured by dynamic landscape features and environments. T. Am. Fish. Soc. 140, 665–684.

Jakobsson M, Rosenberg NA. 2007 CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics 23, 1801–1806.

Matala AP, Ackerman MW, Campbell MR, Narum SR. 2014 Relative contributions of neutral and non-neutral genetic differentiation to inform conservation of steelhead trout across highly variable landscapes. Evol. Appl. 7, 682–701.

Narum SR, Powell MS, Evenson R, Sharp B, Talbot A. 2006 Microsatellites reveal population substructure of Klickitat River native steelhead and genetic divergence from an introduced stock. N. Am. J. Fish. Manage. 26, 147–155.

Narum SR, et al. 2011 Candidate genetic markers associated with anadromy in Oncorhynchus mykiss of the Klickitat River. T. Am. Fish. Soc. 140, 843–854.

Pritchard JK, Stephens M, Donnelly P. 2000 Inference of population structure using multilocus genotype data. Genetics 155, 945–959.

Supplemental Methods S2: Bioinformatics Methods

Samples were run through ustacks with the 'm' parameter based on read count, as follows: < 1.3 M reads, m=2; 1.3 M < x < 2 M, m=3; 2 M < x < 4.5 M, m=5; 4.5 M < x < 7 M, m=6; 7 M < x < 10 M, m=8; and > 10 M, m=10. The cstacks catalog was created with a set of 18 individuals which had been found to have the fewest missing genotypes in preliminary analyses and represented the diversity of migration-timing. The "populations" step was run using m = 5. Using mysql, we only exported tags if they had between 1-4 SNPs per tag (69446). We removed tags that genotyped in ≤50% of samples (26337), excluded monomorphic tags (12) and any tag <1% minor allele frequency (MAF, N=10389), excluded potential PSVs (2360 tags with heterozygote genotypes in 1 or more doubled haploid samples), then excluded poorly genotyped tags (loci that failed across >30% of individuals that genotyped in greater than 70% of all loci, 4719 tags). A whitelist of 25,629 tags was finally examined for MAF (across all 320 individuals) and a single SNP with the highest MAF was chosen per RAD tag, resulting in 18,375 SNPs. (This particular step was critical because an alternative approach to select the first SNP per RAD tag, would have been missed the three most signficant candidate SNPs that were later identified in the GWAS.) Any individuals that failed to genotype at > 75% of the SNPs (n = 83) were removed from further analyses. In order to exclude extremely rare SNPs that might cause spurious associations, the dataset was trimmed further such that the final data set included 15,059 SNPs with > 3% MAF that were successfully genotyped across >80% of 237 individuals. Additional genotypes from 180 TaqMan SNP assays resulted in a total of 15,239 SNPs that were analyzed.

Supplemental Methods S3: Random Forest analyses

For the first approach ("RF-rank"), our "coarse-sweep" involved an iterative process to build a set of predictive models for the migration trait (dependent variable) based on subsets of the 15,239 SNP loci (independent variables). First, an RF analysis with 30,000 trees was performed using all loci, and we found the number of trees should be at least twice the number of loci to provide convergence for ranking the loci based on their importance values (i.e., the relative contribution of each SNP to the RF model's predictive accuracy). Next, the ranking of the markers from the initial RF analysis was used to select a smaller subset of loci (7,500) before performing a subsequent RF analysis. For each RF run, we always used twice the number of trees as the number of loci, and at minimum 2000 trees. After the top 7500 loci RF run completed, we repeated these steps with the subsets of the top 3000, 1500, 750, 500, 400, 300, 200, 100, 75, 50, 25, 10, 5, and 3 loci. This "coarse-sweep" iterative approach of continually reordering loci by importance values from the previous run has been demonstrated to provide substantial improvement versus simply using the same importance values for all subsets of loci based on the initial run (Holliday et al. 2012).

Results from this "coarse-sweep" analysis showed that the maximum phenotypic variance could be explained with the top 25 loci. Using this group of loci, we then ranked their importance using a backward purging analysis (Holliday et al. 2012), which was automated with R scripts (Brieuc et al. 2015). In this backward purging (i.e. "fine sweep" analysis), the least important loci are removed one by one, starting with a greater number than the initial optimum number of loci (here, the best 150 loci rather than the top 25; Supplemental methods). Specifically, the analysis began with the top 150 SNPs from the "coarse-sweep", and three iterations of Random Forest were performed. The locus with the lowest importance value was removed, and another three iterations of Random Forest were performed on the remaining loci.

These steps were repeated until 2 loci remained (the minimum number that can be analyzed with RF), and the results from these steps were used to identify the set of SNPs that explained the greatest amount of trait variation.

References

Brieuc MSO, Ono K, Drinan DP, Naish KA. 2015 Integration of Random Forest with population-based outlier analyses provides insight on the genomic basis and evolution of run timing in Chinook salmon (*Oncorhynchus tshawytscha*). Mol. Ecol. 24, 2729–2746.

Holliday JA, Wang T, Aitken S. 2012 Predicting adaptive phenotypes from multilocus genotypes in Sitka spruce (*Picea sitchensis*) using Random Forest. G3: Genes Genom. Genet. 2, 1085–1093.